SERIEs (2010) 1:459–474
DOI 10.1007/s13209-009-0005-8

ORIGINAL ARTICLE

A queuing-based model for optimal dimension of service firms

Isabel Parra-Frutos

Received: 18 September 2008 / Accepted: 6 July 2009 / Published online: 8 January 2010

Abstract This paper examines a stochastic model to determine optimal pricing, waiting time, output, and sizing decisions for service firms which compete on time in an uncertain environment. Sizing decisions concern optimal service capacity and maximum physical waiting room (with a given probability). Customers are sensitive to money price and expected waiting time. The firm, modeled as an M/M/1 queue, is assumed to be a full price taker as a result of the stochastic version of the perfect competition assumption. Firm costs function includes fixed and variable costs. We initially fix the service rate and obtain a closed-form expression for the waiting time, price, output and waiting room which lead to maximize the net revenue. Subsequently, we obtain first-order conditions for the profit maximizing service capacity. We find that an optimal capacity may exist when service capacity costs are strictly convex. When capacity costs are linear or concave, an optimal capacity does not exist, and it is possible to obtain higher expected profits by increasing service capacity, as the time competition principle states. However, in the linear case we find a failure of that principle, which is, there is a limit in speeding up processes.

Keywords Time-based competition · Service firms · Queuing theory · Firm's capacity

JEL Classification C02 · E10 · M21

I. Parra-Frutos (✉)
Department of Quantitative Methods for Economics and Business, Economics and Business School, University of Murcia, Murcia, Spain
e-mail: ipf@um.es

1 Introduction

In service firms and in make-to-order production firms, waiting time plays a significant role in customers' buying decisions. Customers are required to spend a significant amount of time obtaining the service or product. It was at the end of the 1980s when this aspect was used as a competitive weapon and firms started to compete on delivery time to increase their market shares. Time-based competition emerged as a new competitive weapon that led firms to outstanding success (Stalk 1988; Stalk and Hout 1990; Blackburn 1991; Png and Reitman 1994).

This concern for competition against time attracted special attention in the literature in the 1990s, see, for example Kalai et al. (1992), Loch (1994a, b), Li and Lee (1994), Daniel (1995), Lederer and Li (1997), and Parra-Frutos and Aranda (1999b).

The advantage of using queuing theory to describe service firms lies in the fact that waiting time becomes an endogenous variable, a decision variable of the model. We focus on service firms because their characteristics (direct sales, non-storability, non-transportability, service creation while it is being supplied, etc.) are better represented by a queuing system.

We propose a model, based on an M/M/1 queue, for firms which compete on delivery times and develop their activity in perfect competition. The study is made in terms of expected values for those variables considered random, i.e., in the long run. The aim of the model is to find optimal levels for the decision variables such as service capacity, mean waiting time, money price, expected output rate, and waiting room size (or accumulated orders with a given probability).

The basic idea of the model is that we assume there are some flows, such as those given by demand and service rates, which are clearly related between each other and other variables, for instance, money price, consumer budget, etc. Setting up a business then should be performed so that the resulting adjustment of these flows returns the maximum expected profits. Our model internalizes the corresponding relationships (or internal forces) and the proposed solution describes the final (expected) state of flows and their consequences for money price, waiting time, and waiting room size.

This model belongs to those known as *congestion models*, which are based in the main on queuing theory. They are concerned with optimizing systems where individuals (or elements, in general) suffer or may suffer a delay. In addition, they are characterized by the existence of endogenous externalities among customers, that is, an individual who joins the system causes a negative effect (congestion) on the rest of consumers (Loch 1994a).

It is also a *time competition model*, where customers are sensitive to waiting time, and hence endogenous externalities among customers are also present. Loch (1994a) affirms that time competition models represent a class of *firm differentiation models*, as an alternative to *location models*, where customer externalities are exogenous, i.e., constant and given, and therefore the firm differentiation is exogenous. Differences between this model and those studied in the literature on time competition arise mainly from the control variables, the type of market and the market mechanism. The monopoly market is considered by So and Song (1997) and

Parra-Frutos and Aranda (1999b). The latter considers the possibility of reducing money price when experiencing long waiting times. Duopoly markets are investigated in Kalai et al. (1992), Li and Lee (1994), and Loch (1994a, b). The first studies a queuing system with two servers and only one queue when money price and demand level are given. Li and Lee develop a particular queuing system with two servers and two queues with jockeying. Loch (1994a) uses an M/G/1 queue to describe firms, the consumers are homogeneous, service capacity is given and does not consider costs, and thus, expected incomes are maximized. Loch (1994b) extends the previous model to several customer groups with different levels of impatience, but models firms through an M/M/1 queue. Loch's models consider delivery times endogenous, but they focus on price competition and quantity competition. Oligopolistic markets are examined by Davidson (1982, 1988) and Daniel (1995). Davidson focuses on price advertisement to consider complete and incomplete price information markets. He uses an M/M/1/B queuing model with truncated arrival rate. Demand and service capacity are given. Daniel does not consider firms costs, and the expected service rate is not a control variable for the firms, which are described using an M/M/1 queue where service capacities are identical and given. Perfect competition is considered in Lederer and Li (1997). This model describes firms as an M/G/1 queue which can choose scheduling policies to specify how incoming jobs are sequenced. Firms do not offer homogeneous goods and services and customers are not homogeneous. The model does not consider budget restrictions and fixed costs.

One contribution of our model lies in its optimizing all the firms' decision variables under uncertainty and congestion. There is no variable which firms must fix in advance. In other models, for instance, capacity level or demand is given. Here firms optimize their behavior using consumer knowledge. As Stalk and Webber (1993) state, to be a successful time competitor it is not sufficient to speed up processes but it is also necessary to link consumer knowledge to time-based capabilities. We propose an expected profit function similar to that studied in Daniel (1995), although he neither considers capacity costs nor optimizes service quality nor focuses on an oligopolistic setting.

There are other related studies based on queuing theory, but with a different emphasis. However, none studies time competition. Their interest lies in efficient pricing in a perfect competition environment under uncertainty and congestion. Among them, we can mention, for instance, De Vany and Saving (1977, 1980, 1983) and De Vany et al. (1983). De Vany and Saving (1977) uses the truck transport industry to consider situations where one market dominates the other (two different geographical points). De Vany and Saving (1980) model the competitive supply and pricing of highways with random traffic. The highway system is viewed as a self-service queuing model and the queues that develop at the highway entrance are ignored. De Vany and Saving (1983) use a G/G/1 queue to describe firms with a convex function for costs. De Vany et al. (1983) describe firms as G/G/1 and study some qualitative characteristics of equilibrium without considering the impact of different types of cost functions. They do not take into account the impact of waiting time and service capacity on demand. The reader is referred to Parra-Frutos (1997)

and Parra-Frutos and Aranda (1999a) for a review of this kind of queuing model under different market structures (monopoly, duopoly and perfect competition).

Our concern is to study not only the profit-maximizing firm behavior but also the consumer behavior in a perfect competition environment. In addition, we also focus on a relatively new aspect in competition—time-based competition. In time competition waiting time, as well as price, is a control variable.

The paper is organized as follows. In the next section we describe the model assumptions. In Sect. 3 we investigate a particular profit function to obtain optimal solutions for firms' dimensions. Finally, in Sect. 4 we give some concluding remarks and propose some extensions of the model.

2 The model

2.1 Consumer behavior

The firm offers a service, with money price given by $p > 0$, which can be obtained after waiting a certain time. Since waiting is a fundamental part of a service, customers are not only interested in the money price, but also in the full price. The full price refers to the sum of the money price and waiting time cost.

We consider there is a large number of customers who act independently of each other and base their decisions on their experience and a firm's reputation or publicity rather than on ad hoc search. Customers will take into account *expected* waiting time (W) instead of *actual* waiting time. Consequently, it is the *expected* full price that summarizes all the relevant information for customers. Hence, we are assuming consumers who are sensitive to price and time.

The buying decision rule consists of ordering the service if the expected full price is not higher than the reservation full price (R), which is the highest expected full price consumers are willing to pay

$$p + cW \leq R, \tag{1}$$

where c is the marginal time value of a consumer. We assume that customers are homogenous, so they have the same reservation full price and the same time value.

Customers will only accept the set of combinations of money price and expected waiting time that verifies (1). Put differently, given a money price (expected waiting time) customers will accept an interval of values for the expected waiting time (money price), so that the expected full price is not higher than the reservation full price.

Given a money price and a service capacity, in terms of the expected number of individuals who can be served per time unit, the expected waiting time in a firm is then determined by customers, through the arrival rate. As long as (1) is not verified for equality, consumers have an incentive to place an order with this firm. So, the arrival rate increases until the expected waiting time makes the reservation full price constraint be verified for equality.

There is a tradeoff between money price and waiting time. Customers are willing to wait longer if the money price reduces. So, in order to wait less they accept a higher money price.

2.2 The demand

Customers are assumed to be price and time sensitive. So, the expected arrival rate, λ, depends on the firm's money price and the expected waiting time. So, we have

$$\lambda = \lambda(p, W(\mu)) \tag{2}$$

where μ is the firm's service capacity.

In order to simplify the analysis, we will focus on a particular firm, so we do not need to identify firms with an index. The reader should bear in mind that the analysis is identical for every firm but it does not necessarily imply the same optimal result for the money price, waiting time, and service capacity.

We also assume that customers demand a single service upon their arrival at the firm, and they arrive individually, according to a Poisson process with parameter λ.

2.3 The firm

We assume that the firm's service rate also follows a Poisson process with expected value given by μ. This parameter is also known as the *service capacity* since it refers to the number of customers who may be served per time unit, on average. The service time therefore follows an exponential distribution with expected value $1/\mu$.

We assume that firms fix their service capacity according to market characteristics and their expected profit function. Once customer properties are known and capacity is fixed, firms set their money price so that it leads to an optimum mean waiting time, that is, a profit-maximizing one. Thus, a firm that competes on time uses its decision variables, capacity and money price, to induce a certain mean waiting time.

Firms can be modeled under all these assumptions using an M/M/1 queuing system, whose steady-state solution is (see Gross and Harris 1998)

$$P_k = \rho^k(1-\rho), \quad k = 0, 1, 2, \ldots \tag{3}$$

and

$$\rho = \frac{\lambda}{\mu} < 1, \tag{4}$$

where P_k is the probability of k individuals in the firm waiting or being served. There are different ways of interpreting the parameter ρ (known as the *traffic intensity*). It can be understood as the capacity utilization rate; as the probability of being busy; and finally, as the fraction of each time interval where the firm is busy.

The expected waiting time (queuing time plus service time) is given by (see Gross and Harris 1998)

$$W(\lambda, \mu) = \frac{1}{\mu - \lambda}. \tag{5}$$

Note that an increase in the service capacity, *ceteris paribus*, gives rise to a reduction in the expected waiting time. However, an increase in the arrival rate results in the opposite effect, an increase in system congestion.

We consider that the firm would like at most n individuals in the waiting room, but it is willing to accept more than n with probability equal to or lower than α, where α is very small.

Let N be the random variable "number of individuals in the firm waiting or being served". The firm will establish a waiting room of n individuals so that it verifies

$$\Pr[N > n] \leq \alpha. \tag{6}$$

Lemma 1 *Given λ, μ, and α, the maximum number of individuals (orders) waiting, n, with probability at least $1 - \alpha$* is

$$n \geq \frac{\ln \alpha}{\ln \lambda - \ln \mu} - 1. \tag{7}$$

Proof Using (3) and (6) we obtain

$$\Pr[N > n] = \sum_{i=n+1}^{\infty} \rho^i (1 - \rho) = \rho^{n+1} \leq \alpha. \tag{8}$$

Finally, substituting (4) we can derive (7). □

2.4 The output

It can be shown (see Gross and Harris, 1998, pp. 167–170) that when we are dealing with a queuing system like M/M/c/∞ then the output probability distribution is the same as the input distribution and is not affected at all by the exponential service mechanism. So, if the arrival rate distribution is Poisson with parameter λ then the output also follows a Poisson distribution with identical expected value.

2.5 The Poisson-exponential probability distribution

In the following sections, and in queuing theory in general, the Poisson-exponential distribution plays a key role. There are arguments in favor of it which try to show that it is not a restrictive assumption. One strong argument in favor of exponential inputs is that the limit of a binomial distribution with small probability of success (e.g. entering the system) and defined over a large set of elements or individuals (population) is Poisson. In the second place, one characteristic of a Poisson process is that the moments of time when events occur are uniformly distributed over time (see Ross 1989, p. 224; Law and Kelton 1991, p. 393). Thus, the Poisson-exponential distribution implies that any moment of time is equally likely to have an arrival. Third, there is an additional argument from information theory which says that the exponential distribution provides the least information and is, therefore, the

most random law which can be used and, thus, provides a reasonably conservative approach. The Poisson-exponential distribution therefore represents the most random specification of the stochastic nature of demand and service processes. Relaxing this assumption leads to more deterministic processes (Daniel 1995).

2.6 The stochastic version of the perfect competition assumption

Perfect competition is a market structure mainly characterized by two assumptions: the existence of a high number of firms and consumers and the homogeneity of the product. These assumptions together imply that firms are price-taker or atomistic, i.e., firms cannot influence market prices and have to accept them as given.

The stochastic version of the perfect competition assumption would be the following: firms are takers of the expected full price such that they can sell any level of output at the expected full price of the market. That is, they can receive any expected number of arrivals of customers simply by offering the appropriate level of capacity.

The fact that firms are expected-full-price takers does not imply they are also money-price takers. Indeed, a firm may have any set of money prices, but has to accept the resulting impact on its demand, since the expected waiting time adjusts till the expected full price equals the expected full price of the market.

As customers are indifferent to firms where they place an order in perfect competition, all the firms will offer the same expected full price in equilibrium. This does not imply having the same combination of money price and expected waiting time.

If a firm takes any action which, given its money price, modifies the expected waiting time and makes its expected full price lower than that of the market, then customers will have an incentive to demand this service and therefore will increase their arrival rate until this incentive disappears, i.e., until the expected waiting time is re-established at the previous level—that of other firms (the expected full price of the market). A similar inverse process will start if a firm has an expected full price above that fixed by the market. The consumer behavior will lead to a market expected full price given by R.

Other perfect competition models using queuing theory have considered that the market expected full price is determined somehow by the market (e.g. De Vany and Saving 1980) or as a result of an optimal customer behavior consisting of minimizing expected service costs (Davidson 1988), or as a result of maximizing firm profits (Lederer and Li 1997). These studies did not consider the possibility that the resulting market full price exceeds customer budget restrictions.

3 The expected profit function

We assume a large number of consumers and firms offering a homogeneous service. Firms are greatly influenced by customer characteristics, so these characteristics, in particular, budget constraint and marginal time value, should be taken into account if firms want to optimize their profits.

If we assume that parameters c and R are known by the firm then, given an expected waiting time W, the firm will set that money price which verifies expression (1) on equality. Consequently, to determine the optimum money price, we need to obtain the optimum expected waiting time which leads to a profit-maximizing arrival rate.

We assume that firm costs depend on service capacity and demand. A particular level of service capacity implies a given capability to serve customers, the cost of which has to be paid independently of its level of use. This cost is denoted by $C(\mu)$, where $C(\mu) \neq 0$ if $\mu \neq 0$ and $dC(\mu)/d\mu = C'(\mu) > 0$. In addition, we assume that it is a continuous and derivable function, with the first and second derivatives continuous. On the other hand, there are some variable costs, denoted by γ, that depend on the level of demand, and these are applicable to each service. Thus, γ is a fixed cost per customer.

The demand function (expected arrival rate) is obtained from (5) and the money price function from the reservation expected full price constraint. Hence

$$\lambda(W, \mu) = \mu - \frac{1}{W} \quad \text{where} \quad W > \frac{1}{\mu}, \tag{9}$$

$$p(W) = R - cW \quad \text{where} \quad W < \frac{R}{c}. \tag{10}$$

Conditions for W are derived when imposing $\lambda > 0$ and $p > 0$. The total expected net income is given by the following function

$$I(W, \mu) = \lambda(W, \mu)[p(W) - \gamma]. \tag{11}$$

Hence, the expected profit function for a firm is

$$\begin{aligned} \Pi(W, \mu) &= \text{Total expected net income} - \text{capacity cost} \\ &= \left(\mu - \frac{1}{W}\right)(R - cW - \gamma) - C(\mu), \end{aligned} \tag{12}$$

where $1/\mu < W < R/c$. In this formulation it can be observed that W and μ are the decision variables. Consequently, the firm has to install a service capacity level and set the money price that induces the waiting time desired. The remaining variables, output and waiting room size, will adjust according to the market mechanism summarized in (9) and (7).

Before studying the characteristics of the expected profit function (which we group in Lemma 2) we investigate the *conditions* on μ and W such that $\Pi(W, \mu)$ is positive. If it is null or negative, then the firm will decide not to set up in that market, or will leave if it is already operating in it. Therefore, the maximization problem would not make sense.

Condition A A necessary condition on W is,

$$\frac{1}{\mu} < W < \frac{R - \gamma}{c}. \tag{13}$$

This condition gives the values of W that make the total expected net income positive. This happens when the arrival rate $\lambda(W, \mu)$ and the net income per customer

$p(W) - \gamma$ are, in turn, positive. The first inequality is derived using (9). Note that if the maximum mean waiting time accepted by individuals is equal or inferior to the mean service time of the firm $(1/\mu)$, the firm will have no demand and, hence, a loss equal to the capacity cost. Second, to have a positive net income per customer, it must also occur that $p(W) > \gamma$, and using (10) we obtain $W < (R - \gamma)/c$.

This condition is implicitly including a restriction on p given by

$$\gamma < p < R. \tag{14}$$

The first part of this inequality has been used to obtain (13). It can be derived that p is lower than R by using in (10) the upper bound for W given in (13).

Condition B The following condition on service capacity can be derived from (13), since $1/\mu < (R - \gamma)/c$ must occur,

$$\mu > \frac{c}{R - \gamma}. \tag{15}$$

This condition gives the possible range of values for service capacity according to consumer characteristics. Note that if condition A verifies, then condition B verifies as well, but not vice versa. For the sake of simplicity we denote the minimum service capacity level $c/(R - \gamma)$ as μ^*.

Condition C The capacity cost function is such that the expected profit function is positive for some W and μ.

Therefore, the firm's expected profit function may be finally written as

$$\Pi(W, \mu) = \begin{cases} \left(\mu - \frac{1}{W}\right)(R - cW - \gamma) - C(\mu) & \text{if} \quad W \in \left(\frac{1}{\mu}, \frac{R-\gamma}{c}\right) \quad \text{and} \quad \mu > \mu^* \\ -C(\mu) & \text{rest} \end{cases} \tag{16}$$

where conditions A and B have been included. Finally, we impose that condition C must also be verified. Before studying optimal solutions for the problem of maximizing profits we investigate the properties of this function in the following lemma.

Lemma 2 *The expected profit function* $\Pi(W, \mu)$ *has the following properties*:

(i) *For any fixed value* $\mu = \mu_0$, $\Pi(W; \mu_0)$ *is a continuous function in* $W>0$, *whose first and second derivatives are continuous in* $W \in (1/\mu_0, (R - \gamma)/c)$.

(ii) *For any fixed value* $\mu = \mu_0$, $\Pi(W; \mu_0)$ *is a strictly concave function with respect to* W, $\forall W > 0$.

(iii) *There exist two values* W_1 *and* W_2, *where* $1/\mu < W_1 < W_2 < (R - \gamma)/c$, *such that* $\Pi(W, \mu) > 0$ *if* $W \in (W_1, W_2)$.

(iv) *There exists a maximum in* $\Pi(W, \mu)$ *when* $W \in (1/\mu, (R - \gamma)/c)$.

(v) *For any fixed value* $W = W_0$, *such that* $0 < W_0 < (R - \gamma)/c$, $\Pi(\mu; W_0)$ *is a continuous function of* μ, *where* $\mu > \mu^*$.

(vi) *For any fixed value $W = W_0$, such that $0 < W_0 < (R - \gamma)/c$, $\Pi(\mu; W_0)$ is a convex (concave) function with respect to μ if $C(\mu)$ is concave (convex). If a minimum (maximum) exists, then it verifies $p - \gamma = C'(\mu)$. In addition, it is monotonic increasing for those values of μ such that $p - \gamma > C'(\mu)$, and monotonic decreasing when the contrary is the case.*

Proof To prove property (i) note that possible discontinuities may be at $W = 1/\mu_0$ and $W = (R - \gamma)/c$. However,

$$\lim_{W \to 1/\mu_0^-} \Pi(W; \mu_0) = \lim_{W \to 1/\mu_0^+} \Pi(W; \mu_0) = \Pi\left(1/\mu_0, \mu_0\right) = -C(\mu_0), \tag{17}$$

$$\lim_{W \to (R-\gamma)/c^-} \Pi(W; \mu_0) = \lim_{W \to (R-\gamma)/c^+} \Pi(W; \mu_0) = \Pi((R - \gamma)/c, \mu_0) = -C(\mu_0), \tag{18}$$

Hence, $\Pi(W; \mu_0)$ is continuous in $W > 0$. Its first and second derivatives are

$$\frac{\partial \Pi(W; \mu_0)}{\partial W} = \frac{R - \gamma}{W^2} - c\mu_0, \tag{19}$$

$$\frac{\partial^2 \Pi(W; \mu_0)}{\partial W^2} = \frac{-2(R - \gamma)}{W^3}. \tag{20}$$

It can be easily shown that both functions are continuous in $1/\mu_0 < W < (R - \gamma)/c$. Property (ii) is easily shown since (20) is negative, given that it verifies expression (14) and $W > 0$.

To show property (iii) we have, using (16), that $\Pi(W, \mu) < 0$ at $W = 1/\mu$ and $W = (R - \gamma)/c$. Taking into account conditions A and C, we can affirm that there exists an open interval $(W_1, W_2) \subset (1/\mu, (R - \gamma)/c)$ where $\Pi(W, \mu)$ is positive. In addition, from *Bolzano's theorem*, $\Pi(W_1, \mu) = \Pi(W_2, \mu) = 0$.

Property (iv) derives from (i), (ii), and (iii). From property (i) $\Pi(W, \mu)$ is continuous; from property (ii) it is strictly concave; and finally from property (iii) $\Pi(W, \mu) < 0$ at $W = 1/\mu$ and $W = (R - \gamma)/c$, and $\Pi(W, \mu) > 0$ at $(W_1, W_2) \subset (1/\mu, (R - \gamma)/c)$. Therefore, $\Pi(W, \mu)$ has a maximum at $\hat{W} \in (W_1, W_2)$ such that $\partial \Pi / \partial W\big|_{W=\hat{W}} = 0$.

To prove property (v) we should take into account the assumption of continuity of $C(\mu)$.

Property (vi) is derived by calculating the corresponding derivatives of $\Pi(\mu; W_0)$ with respect to μ

$$\frac{\partial \Pi(\mu; W_0)}{\partial \mu} = R - cW_0 - \gamma - C'(\mu); \tag{21}$$

$$\frac{\partial^2 \Pi(\mu; W_0)}{\partial \mu^2} = -C''(\mu) \tag{22}$$

□

3.1 Optimal solution for firms already installed in the market

The optimal solution derived from maximizing expected profits with respect to W when the service capacity is already installed is given in the following theorem.

Theorem 1 *Let a service firm be a time competitor in a market in perfect competition modeled by a queuing system M/M/1 whose demand is composed of one class of customer with a reservation full price R and marginal time cost c. Then, to have a profit-maximizing waiting time given a service capacity installed μ_I, such that $\mu_I > \mu^*$, a firm should fix the money price.*

$$\hat{p} = R - \sqrt{\frac{(R-\gamma)c}{\mu_I}}, \tag{23}$$

The resulting mean waiting time is, then, the optimum one for the firm given by

$$\hat{W} = \sqrt{\frac{R-\gamma}{\mu_I c}}. \tag{24}$$

The optimum expected output rate is then

$$\hat{\lambda} = \mu_I - \sqrt{\frac{\mu_I c}{R-\gamma}}, \tag{25}$$

where $\hat{\lambda}$ is strictly increasing in μ_I. Finally, the optimum waiting room size with probability of at least $1 - \alpha$ is

$$\hat{n} \geq \frac{\ln \alpha}{\ln \hat{\lambda} - \ln \mu_I} - 1 \quad 0 < \alpha < 1. \tag{26}$$

Proof From property (iv) of Lemma 2 the expected profit function (16) presents a maximum in the interval $(1/\mu, (R-\gamma)/c)$. The first-order condition to maximize (16) allows us to derive an expression for the optimum mean waiting time for a service capacity $\mu = \mu_I$ and is given, using (19), by

$$\hat{W} = \sqrt{\frac{R-\gamma}{\mu_I c}}, \tag{27}$$

Using (10), the appropriate money price is

$$\hat{p} = R - c\hat{W}.$$

From (9) and (27) the expected output rate is given by

$$\hat{\lambda} = \mu_I - \sqrt{\frac{\mu_I c}{R-\gamma}},$$

where

$$\frac{\partial \hat{\lambda}}{\partial \mu_I} = 1 - \frac{1}{2}\sqrt{\frac{c}{R-\gamma}}\mu_I^{-1/2}$$

$$\frac{\partial^2 \hat{\lambda}}{\partial \mu_I^2} = \frac{1}{4}\sqrt{\frac{c}{R-\gamma}}\mu_I^{-3/2} > 0.$$

From here, it can be derived that $\hat{\lambda}$ is convex and has a minimum at $\tilde{\mu} = c/[4(R-\gamma)]$, where $\tilde{\mu} < \mu^*$. Therefore, $\hat{\lambda}$ is strictly increasing with respect to μ, $\forall \mu > \mu^*$, such that the stochastic version of the perfect competition assumption is verified.

Finally, using Lemma 1, for a small α (where $0 < \alpha < 1$) fixed by the firm, the optimum waiting room size is

$$\hat{n} \geq \frac{\ln \alpha}{\ln \hat{\lambda} - \ln \mu_I} - 1.$$

□

In this section we assume that the firm is already installed in the market with an appropriate service capacity level. The firm decides to compete on time and then fix a money price that leads to a profit-maximizing waiting time. The resulting output (the expected number of individuals the firm should serve per unit time) and waiting room size (if physical presence is required, otherwise this is equivalent to the maximum number of accumulated orders with probability at least $1 - \alpha$) are given in (25) and (26), respectively.

3.2 Optimal dimension of service firms

If the service firm is not already installed or it is possible to adjust its service capacity, then it should fix an optimum level. In this section we investigate the expected profit at the optimum mean waiting time as a function of service capacity, that is, the *optimal expected profit function* (OEPF), which is denoted by $\Pi^*(\mu)$. We use this function to derive those values of the service capacity which give the highest level of optimal expected profit.

Substituting (24) for $\forall \mu$ in (16) we obtain the OEPF $\Pi^*(\mu)$

$$\Pi^*(\mu) = \Pi\big(\mu; \hat{W}\big) = (R-\gamma)\mu - 2\sqrt{c(R-\gamma)}\sqrt{\mu} + c - C(\mu), \tag{28}$$

where $\mu > \mu^*$.

Observe that $\Pi^*(\mu^*) = -C(\mu^*) < 0$. Note also that $\Pi^*(\mu)$ is different from $\Pi(\mu; W_0)$ studied in property (vi) of Lemma 2. The second function is defined for a fixed value $W = W_0$, and the first one allows W to vary.

The first-order condition to maximize (28) leads to

$$C'(\hat{\mu}) + \sqrt{c(R-\gamma)}\hat{\mu}^{-1/2} = R - \gamma, \tag{29}$$

where $\hat{\mu}$ may not exist, depending on $C(\mu)$. Moreover, if $\hat{\mu}$ exists, an explicit expression may not be always found. The second-order condition is

$$\frac{\partial^2 \Pi^*(\mu)}{\partial \mu^2} = \frac{1}{2}\sqrt{c(R-\gamma)}\mu^{-3/2} - C''(\mu). \tag{30}$$

The results of the study of $\hat{\mu}$, if it exists, are given in the following lemma.

Lemma 3 *If $\hat{\mu}$ exists, then $\hat{\mu} > \mu^*$.*

Proof Since $C'(\mu) > 0$, using (29) we have

$$\sqrt{c(R-\gamma)}\hat{\mu}^{-1/2} < R - \gamma.$$

Thus,

$$\hat{\mu} > \frac{c}{R-\gamma} = \mu^*.$$

□

This lemma affirms that if $\hat{\mu}$ exists and optimizes $\Pi^*(\mu)$ when $\mu \in R^+$, then it also belongs to the domain of $\Pi^*(\mu)$ given by $\mu > \mu^*$. Using (30), $\Pi^*(\mu)$ may be concave or convex depending on the functional form of capacity cost $C(\mu)$. Consequently, $\hat{\mu}$, if it exists, will be a maximum or a minimum.

We now investigate the consequences, for the OEPF and service capacity decision, of different functional forms for the capacity cost. In particular, we discuss a linear, concave and convex capacity cost function, that is, when $C''(\mu) = 0$, $C''(\mu) = k_1 < 0$ (with k_1 constant), and $C''(\mu) = k_2 > 0$ (with k_2 constant), respectively.

Lemma 4 *$\Pi^*(\mu)$ is convex if $C(\mu)$ is linear or concave, and it may present a minimum in $\hat{\mu}$, if $\hat{\mu}$ exists (where $\Pi^*(\hat{\mu}) < 0$), or be strictly increasing if it does not exist.*

Proof If $C(\mu)$ is linear, then $C''(\mu) = 0$. Thus, the second-order condition (30) is positive, indicating that $\Pi^*(\mu)$ is convex. If $C(\mu)$ is concave then $C''(\mu) < 0$, hence the second-order condition (30) is also positive and, consequently, $\Pi^*(\mu)$ is again convex. Finally, if $\hat{\mu}$ exists, then $\Pi^*(\hat{\mu}) < 0$, since $\Pi^*(\mu^*) < 0$ and $\hat{\mu}$ is a minimum. If $\hat{\mu}$ does not exist then $\Pi^*(\mu)$ is strictly increasing since $\Pi^*(\mu^*) < 0$ and there is some μ such that $\Pi^*(\mu) > 0$ (from condition D). □

Consequently, if $C(\mu)$ is linear or concave, the optimal dimension of a service firm is given by the following theorem.

Theorem 2 *If $C(\mu)$ is linear or concave, then there exists a capacity level μ^{**}, such that $\mu^{**} > \mu^*$ and $\Pi^*(\mu^{**}) = 0$, from which optimal expected profits are positive and increasing in μ.*

Proof Recall that $\Pi^*(\mu^*) < 0$ and, from Condition C, expected profits are positive for some μ. If $C(\mu)$ is linear or concave, then $\Pi^*(\mu)$ is convex from Lemma 4. It follows that there exists a capacity level μ^{**} such that $\Pi^*(\mu^{**}) = 0$, and $\Pi^*(\mu) > 0 \ \forall \mu > \mu^{**}$. Thus, the higher the service capacity level, the higher the optimal expected profits. □

Consequently, if $C(\mu)$ is linear or concave, then there is no optimal dimension of firms, i.e., a dimension which gives rise to the maximum level possible in the expected profit function. It is always possible to obtain a higher level of expected profits merely by increasing service capacity.

Bear in mind that an increase in service capacity, *ceteris paribus*, results in a lower expected waiting time. So, when $C(\mu)$ is linear or concave, if a firm engages in time competition it may speed up its processes, thus increasing its service capacity. Consequently, it will enlarge its expected profits.

Lemma 5 *If $C(\mu)$ is convex then $\Pi^*(\mu)$ is concave and may present a first convex span if*

$$C''(\mu) < \frac{1}{2}\sqrt{c(R-\gamma)}\mu^{-3/2}. \tag{31}$$

In addition, a minimum (maximum) may exist in the convex (concave) span.

Proof $C(\mu)$ is convex then $C''(\mu) = k_2 > 0$. On the other hand, $1/2\sqrt{c(R-\gamma)}\mu^{-3/2}$ is positive and strictly decreasing in μ. Thus, the second-order derivative (30) may be positive, for small μ, and hence $\Pi^*(\mu)$ could firstly be convex and later become concave as μ increases. □

As a summary of summarize the previous results, and taking into account that $\Pi^*(\mu^*) < 0$ we give the following corollary.

Corollary 1 *If $C(\mu)$ is convex then the supreme value of $\Pi^*(\mu)$ is located in its concave span.*

We have not considered a limit on a firm's demand. This may be valid, for example, in markets in continuous expansion. However, it may also be thought that such a limit exists and is given by the total expected market demand (per time unit), which coincides with the sum of each firm's demand. On the other hand, this limit may be considered variable and dependent, for instance, on time (due to increases or decreases in population), marketing expenses, supply level (due to the fact that a higher supply gives rise to a higher or lower demand), etc. Note that the analysis performed is long term since we are working with expected values.

The existence of such a limit for demand would imply that $\hat{\lambda}$ is not strictly increasing $\forall \mu > \mu^*$, as we obtained in Theorem 1, but monotonic non-decreasing. In other words, demand function is constant from a service capacity level, say K, for which total market demand is achieved. To fix a capacity level above this level gives rise to a reduction in waiting time, since demand does not increase. From the firm's point of view, it would imply higher capacity costs for the same demand level and, thus, firm profits decrease.

Consequently, using Theorem 2 and Corollary 1, and the possible existence of a limit for the service capacity, we may conclude the following:

(a) When C(μ) is linear or concave, the optimum capacity level is given by the highest possible one, K. However, a different issue is that the firm may serve the whole market with only one server. Therefore, there may exist another

limit lower than K, which would be determined, for instance, by technology or by physical limitations of the server. In this way, the optimum firm solution is to serve at its highest rate possible.

(b) When $C(\mu)$ is convex then the optimum capacity level may coincide with $\hat{\mu}$, which maximizes OEPF. Obviously, this depends on the value of K. That is, if $K < \hat{\mu}$ then K is optimum level of μ, otherwise the optimum value is $\hat{\mu}$. Furthermore, it may occur, as in the previous case, that K is an unattainable value for a single server, since there exists a technical or physical limit for the service capacity which prevents a single server from supplying the whole market and thus being a monopoly.

Note that we found that if $C(\mu)$ is convex [$\Pi^*(\mu)$ is concave] and we can assume no limit for capacity level, the firm may have an optimal service capacity level such that optimal expected profits are maximized. In this case, we can affirm that an increase in service capacity (and hence a time competition) does not always lead to higher expected profits. This result contradicts what has generally been asserted as an "advantage of time competition". In other words, we find that in some cases this advantage may have a limit.

4 Concluding remarks

The model studied here presents results on firms' equilibria when competing against time. We show that the strategy firms must follow differs depending on the functional form of service capacity cost. When service capacity cost is a linear or concave function, there is no optimal solution for firms, since the higher the capacity, the higher the expected profits they can obtain. When capacity cost is convex there may be a unique optimal solution. Accordingly, in certain cases the principle of time competition, which affirms that speeding up processes leads to higher profits, cannot be applied.

It should be pointed out that the optimal solution, in the case that it exists, does not have to be the same for every firm in the market. It would be the same only if firms have the same cost structure. Therefore, the market may be characterized by firms offering different combinations of money price and mean waiting time.

We assumed that marginal time cost is linear. This may be seen as a restrictive assumption since it implies no psychological cost of waiting (e.g. the first minute of waiting has the same value of the nth). It is generally accepted that waiting produces negative feelings in individuals, such as frustration, boredom, anxiety, stress, etc. Thus, there is a psychological cost associated with waiting which we did not take into account. Unfortunately, much of the voluminous research to date on queuing has not considered it. Indeed, it is a field of study in which surprisingly little research has been conducted.

It is also possible to expand the model to consider different classes of customers, that is, assuming that customers are heterogeneous and have different budget constraints and marginal time value. This would allow the study of how the market

accommodates customers with different time preferences and, therefore, the nature of equilibrium.

References

Blackburn JD (1991) Time-based competition. Richard Irwin, Homewood
Daniel JI (1995) Price and waiting time dispersion among service firms. Working paper, College of Business and Economics, University of Delaware, Newark
Davidson C (1982) Equilibrium in oligopolistic service industries: an economic application of queuing theory. Social systems research institute workshop series paper no. 8217. University of Wisconsin-Madison
Davidson C (1988) Equilibrium in service industries: an economic application of queuing theory. J Bus 61:347–367
De Vany AS, Saving TR (1977) Product quality, uncertainty and regulation: the trucking industry. Am Econ Rev 67:583–594
De Vany AS, Saving TR (1980) Competition and highway pricing for stochastic traffic. J Bus 53:45–60
De Vany AS, Saving TR (1983) The economics of quality. J Polit Econ 91:979–1000
De Vany AS, Gramm WL, Saving TR, Smithson CW (1983) Production in a service industry using customer inputs: a stochastic model. Rev Econ Stat 65:149–153
Gross D, Harris CM (1998) Fundamentals of queueing theory, 3rd edn (1st edn. 1974, and 2nd edn. 1985). Wiley, New York
Kalai E, Kamien MI, Rubinovitch M (1992) Optimal service speeds in a competitive environment. Manage Sci 38:1154–1163
Law AM, Kelton WD (1991) Simulation modeling and analysis, 2nd edn. McGraw-Hill, New York
Lederer PJ, Li L (1997) Pricing, production, scheduling, and delivery-time competition. Oper Res 45:407–420
Li L, Lee YS (1994) Pricing and delivery-time performance in a competitive environment. Manage Sci 40:633–646
Loch C (1994a) Time competition is capability competition. INSEAD Working paper series
Loch C (1994b) Incentive compatible equilibria in markets with time competition. INSEAD working paper series
Parra-Frutos I (1997) Estructuras de Mercado Tradicionales bajo la Teoría de Colas, Trabajo de Investigación de Tercer Ciclo (Research Dissertation), Departamento de Métodos Cuantitativos para la Economía, Universidad de Murcia
Parra-Frutos I, Aranda J (1999a) Modelos de Mercado: Una aplicación de la Teoría de Colas. Estudios de Economía Aplicada 11:121–142
Parra-Frutos I, Aranda J (1999b) Multiproduct monopoly: a queuing approach. Appl Econ 31:567–578
Png IPL, Reitman D (1994) Service time competition. Rand J Econ 25:619–634
Ross SM (1989) Introduction to probability models, 4th edn. Academic Press, San Diego
So KC, Song J (1997) Price, delivery time guarantees, and capacity selection. Working paper, Graduate School of Management, University of California, Irvine
Stalk G Jr (1988) Time-the next source of competitive advantage. Harvard Bus Rev July–August:41–51
Stalk G Jr, Hout TM (1990) Competing against time. The Free Press, New York
Stalk G Jr, Webber AM (1993) Japan's dark side of time. Harvard Bus Rev July–August:93–102